Title: Challenges of COVID-19 Case Forecasting in the US, 2020-2021

Authors: Velma K Lopez [1]*, Estee Y Cramer [2], Robert Pagano [3], John M Drake [4], Eamon B O'Dea [4], Madeline Adee [5], Turgay Ayer [6], Jagpreet Chhatwal [7], Ozden O Dalgic [8], Mary A Ladd [5], Benjamin P Linas [9], Peter P Mueller [7], Jade Xiao [6], Johannes Bracher [10], Alvaro J Castro Rivadeneira [2], Aaron Gerding [2], Tilmann Gneiting [11], Yuxin Huang [2], Dasuni Jayawardena [2], Abdul H Kanji [2], Khoa Le [2], Anja Mühlemann [12], Jarad Niemi [13], Evan L Ray [2], Ariane Stark [2], Yijin Wang [2], Nutcha Wattanachit [2], Martha W Zorn [2], Sen Pei [14], Jeffrey Shaman [14], Teresa K Yamana [14], Samuel R Tarasewicz [15], Daniel J Wilson [15], Sid Baccam [16], Heidi Gurung [16], Steve Stage [16], Brad Suchoski [16], Lei Gao [17], Zhiling Gu [13], Myungjin Kim [18], Xinyi Li [19], Guannan Wang [20], Lily Wang [17], Yueying Wang [21], Shan Yu [22], Lauren Gardner [23], Sonia Jindal [23], Maximilian Marshall [23], Kristen Nixon [23], Juan Dent [24], Alison L Hill [23], Joshua Kaminsky [24], Elizabeth C Lee [24], Joseph C Lemaitre [25], Justin Lessler [26], Claire P Smith [24], Shaun Truelove [24], Matt Kinsey [27], Luke C. Mullany [27], Kaitlin Rainwater-Lovett [27], Lauren Shin [27], Katharine Tallaksen [27], Shelby Wilson [27], Dean Karlen [28], Lauren Castro [29], Geoffrey Fairchild [29], Isaac Michaud [29], Dave Osthus [29], Jiang Bian [30], Wei Cao [30], Zhifeng Gao [30], Juan Lavista Ferres [30], Chaozhuo Li [30], Tie-Yan Liu [30], Xing Xie [30], Shun Zhang [30], Shun Zheng [30], Matteo Chinazzi [31], Jessica T Davis [31], Kunpeng Mu [31], Ana Pastore y Piontti [31], Alessandro Vespignani [31], Xinyue Xiong [31], Robert Walraven [32], Jinghui Chen [33], Quanquan Gu [33], Lingxiao Wang [33], Pan Xu [33], Weitong Zhang [33], Difan Zou [33], Graham Casey Gibson [34], Daniel Sheldon [2], Ajitesh Srivastava [35], Aniruddha Adiga [22], Benjamin Hurt [22], Gursharn Kaur [22], Bryan Lewis [22], Madhav Marathe [22], Akhil Sai Peddireddy [36], Przemyslaw Porebski [22], Srinivasan Venkatramanan [22], Lijing Wang [37], Pragati V Prasad [1], Jo W Walker [1], Alexander E Webber [1], Rachel B Slayton [1], Matthew Biggerstaff [1], Nicholas G Reich [2], Michael Johansson [1]

Affiliations:

[1] COVID-19 Response, Centers for Disease Control and Prevention, Atlanta, GA, United States of America
[2] University of Massachusetts, Amherst, Amherst, MA, United States of America
[3] Unaffiliated, Tucson, AZ, United States of America
[4] University of Georgia, Athens, GA, United States of America
[5] Massachusetts General Hospital, Boston, MA, United States of America
[6] Georgia Institute of Technology, Atlanta, GA, United States of America
[7] Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States of America
[8] Value Analytics Labs, Boston, MA, United States of America
[9] Boston University School of Medicine, Boston, MA, United States of America
[10] Chair of Econometrics and Statistics, Karlsruhe Institute of Technology, Karlsruhe, Germany
[11] Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
[12] Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland
[13] Iowa State University, Ames, IA, United States of America
[14] Mailman School of Public Health, Columbia University, New York, New York, United States of America
[15] Federal Reserve Bank of San Francisco, San Francisco, CA, United States of America
[16] IEM, Bel Air, MD, United States of America
[16] IEM, Baton Rouge, LA, United States of America
[17] George Mason University, Fairfax, VA, United States of America
[18] Kyungpook National University, Bukgu, Daegu, Republic of Korea
[19] Clemson University, Clemson, SC, United States of America
[20] College of William & Mary, Williamsburg, VA, United States of America
[21] Amazon, Seattle, WA, United States of America

[22] University of Virginia, Charlottesville, VA, United States of America
[23] Johns Hopkins University, Baltimore, MD, United States of America
[24] Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States of America
[25] École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[26] UNC Gillings School of Global Public Health, Chapel Hill, NC, United States of America
[27] Johns Hopkins University Applied Physics Lab, Baltimore, MD, United States of America
[28] University of Victoria, Victoria, BC, Canada
[29] Los Alamos National Laboratory, Los Alamos, NM, United States of America
[30] Microsoft, Redmond, WA, United States of America
[31] Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston, MA, United States of America
[32] Unaffiliated, Davis, CA, United States of America
[33] University of California, Los Angeles, Los Angeles, CA, United States of America
[34] Los Alamos National Lab, Los Alamos, NM, United States of America
[35] University of Southern California, Los Angeles, CA, United States of America
[36] Discreet Labs, Raleigh, NC, United States of America
[37] New Jersey Institute of Technology, Newark, NJ, United States of America

**Corresponding author: oko8@cdc.gov (VKL)*

Abstract

During the COVID-19 pandemic, forecasting COVID-19 trends to support planning and response was a priority for scientists and decision makers alike. In the United States, COVID-19 forecasting was coordinated by a large group of universities, companies, and government entities led by the Centers for Disease Control and Prevention and the US COVID-19 Forecast Hub (https://covid19forecasthub.org). We evaluated approximately 9.7 million forecasts of weekly state-level COVID-19 cases for predictions 1-4 weeks into the future submitted by 24 teams from August 2020 to December 2021. We assessed coverage of central prediction intervals and weighted interval scores (WIS), adjusting for missing forecasts relative to a baseline forecast, and used a Gaussian generalized estimating equation (GEE) model to evaluate differences in skill across epidemic phases that were defined by the effective reproduction number. Overall, we found high variation in skill across individual models, with ensemble-based forecasts outperforming other approaches. Forecast skill relative to the baseline was generally higher for larger jurisdictions (e.g., states compared to counties). Over time, forecasts generally performed worst in periods of rapid changes in reported cases (either in increasing or decreasing epidemic phases) with 95% prediction interval coverage dropping below 50% during the growth phases of the winter 2020, Delta, and Omicron waves. Ideally, case forecasts could serve as a leading indicator of changes in transmission dynamics. However, while most COVID-19 case forecasts outperformed a naïve baseline model, even the most accurate case forecasts were unreliable in key phases. Further research could improve forecasts of leading indicators, like COVID-19 cases, by leveraging additional real-time data, addressing performance across phases, improving the characterization of forecast confidence, and ensuring that forecasts were coherent across spatial scales. In the meantime, it is critical for forecast users to appreciate current limitations and use a broad set of indicators to inform pandemic-related decision making.

Author Summary

As SARS-CoV-2 began to spread throughout the world in early 2020, modelers played a critical role in predicting how the epidemic could take shape. Short-term forecasts of epidemic outcomes (for example, infections, cases, hospitalizations, or deaths) provided useful information to support pandemic planning, resource allocation, and intervention . Yet, infectious disease forecasting is still a nascent science, and the reliability of different types of forecasts is unclear. We retrospectively evaluated COVID-19 case forecasts, which were often unreliable. For example, forecasts did not anticipate the speed of increase in cases in early winter 2020. This analysis provides insights on specific problems that could be addressed in future research to improve forecasts and their use. Identifying the strengths and weaknesses of forecasts is critical to improving forecasting for current and future public health responses.

Introduction

Predicting the trajectory of an epidemic to support control and mitigation planning is the primary objective of infectious disease forecasting. To this end, large-scale, collaborative forecasting efforts across multiple disease systems, such as influenza (1–3), dengue (4), West Nile (5), and Ebola viruses (6), have been integrated into routine public health workflows and emergency response (7). Researchers in academia, private institutions, and the United States (US) government built upon these frameworks to incorporate forecasts into the COVID-19 information systems used to inform pandemic response and created the US COVID-19 Forecast Hub. In April 2020, the US Centers for Disease Control and Prevention (CDC) and the COVID-19 Forecast Hub began collecting COVID-19 death forecasts (8). Compared to death reports, case reports are a leading indicator of SARS-CoV-2 infections, as the time from infection to case report is typically shorter than that between infection and death report. Hence, information gleaned from case forecasts is potentially more actionable.

Case forecasts for all US counties (n=3,143), states (n=50), territories (n=5), the District of Columbia (DC), and the nation as a whole were generated and collected beginning in July 2020, with ensemble forecasts of cases first posted on a CDC webpage on August 6, 2020 (8,9). Because of their potential utility, case forecasts were also integrated into US government web pages and situational awareness updates (10). In addition, county-level case forecasts were used to inform vaccine trial site selection (11) and COVID-19 case forecasts have been cited as useful for guiding personal risk-based decisions (12). Because these forecasts influence policies and personal decisions, accuracy and precision of the forecasts is of the utmost importance. Incorrect forecasts can lead to inappropriate policy implementation and resource allocation, and also to erosion of trust in public health institutions (13).

As part of routine use of the case forecasts in the COVID-19 response, real-time evaluation was conducted. One of the performance metrics included in the evaluation was the 95% prediction interval (PI) coverage, an estimate of the frequency at which the interval captures the eventually observed data. The 95% PI of a reliable forecast should capture eventually reported cases 95% of the time. However, the real-time evaluation indicated that case forecasts were not always reliable, with much lower 95% PI coverage than expected (14). For example, in November 2020 as the 2020-2021 winter wave began, the 95% PI coverage for all states and territories was less than 50% for even the shortest, 1-week ahead forecasts from the ensemble – generally the most reliable forecast. Repeated periods of low coverage during subsequent surges led CDC to stop posting COVID-19 case forecasts in December 2021. Though these forecasts showed poor performance, there are opportunities to develop more precise and reliable future predictions.

Evaluation of forecast performance provides an opportunity not only to assess prediction skill for the purposes of improving forecasts, but also to assess the reliability of the forecasts and foster transparency between forecast users and creators. While evaluation is recommended in forecasting research guidelines (i.e., EPIFORGE 2020 (15), a systematic review of COVID-19 models showed that half of published models did not include probabilistic predictions and that approximately one-fourth of published models did not include performance evaluations (16). We have previously evaluated forecast performance of cumulative (17) and incident (18) COVID-19 deaths submitted to the COVID-19 Forecast Hub. Given that an ensemble of submitted models provided consistently accurate probabilistic forecasts at different scales in both evaluations, here we apply similar methods to assess the prediction skill of the COVID-19 case forecasters, with particular interest in the COVIDhub ensemble model (that is, a model that combine predictions from forecasts submitted to the Forecast Hub). Specifically, we analyze prediction interval coverage and other aspects of nearly 10 million individual forecasts collected by the COVID-19 Forecast Hub for US jurisdictions between July 2020 and December 2021, the full period over which COVID-19 case forecasts were published by the CDC. We analyze relative forecast performance across spatial scales and phases of the pandemic to identify limitations and opportunities for future improvement of case forecasts.

Results

Summary of Included Team Forecasts

A total of 14,960,171 forecasts were submitted by 67 teams throughout the analysis period (see Supporting Information [S] 1 for submission patterns over time). Because forecasts were submitted at multiple geographic scales, we stratified analyses for 1) national forecasts, 2) state (including all 50 states), territory (US Virgin Islands and Puerto Rico), and DC forecasts), 3) county level forecasts (include all 3,143 counties and county equivalents), split into five equal sized groups based on county population size.

We first evaluated forecasts for inclusion criteria based on numbers of locations, horizons, and time periods forecast with the same model. Briefly, teams were included if they submitted the full range of required quantiles, included at least 50 of states/territories/DC or 75% of counties, and produced forecasts at least four weeks into the future for at least 50% of the time points in the study period. At the national level, 22 sets of team forecasts met these criteria (5,136 forecasts across dates and forecast horizons), 23 sets of team forecasts met the state/territory level criteria (280,132 forecasts across jurisdictions, dates, and forecast horizons), and 15 sets of team forecasts met the county-level criteria (9,415,460 forecasts across counties, dates, and forecast horizons). Overall, 64.8% of all submitted forecasts were included in the analysis (9,700,728 forecasts). Of the included forecasts, 11 sets of team forecasts met the inclusion criteria for analyzing submissions across all geospatial scales (8,125,220 forecasts for specific locations, date and forecast horizon).

Each team included in the analysis submitted forecasts that were generated from unique model structures, data inputs, and assumptions (S1). Two naïve models (the COVIDhub-baseline and CEID-Walk) and four ensemble models (the COVIDhub-4_week_ensemble, the COVIDhub-trained_ensemble, LNQ-ens1, and UVA-Ensemble), which combined multiple forecasts into one, were included in the 26 models evaluated (see S1 Table 1.1). The COVIDhub-baseline model projects the number of reported cases in the most recent week as the median predicted value for the next 4 weeks. CEID-Walk is a random walk model with a simple method for removing outliers. A total of seven models included data on COVID-19 hospitalizations, 12 models incorporated demographic data, and seven models used mobility data. Of the 26 evaluated models, three assumed that social distancing and other behavioral patterns changed during the prediction period.

The evaluation period consisted of 1-4 week ahead forecasts submitted in the 73 weeks from July 28, 2020 through December 21, 2021. Multiple phases of the US epidemic were included: the late summer 2020 increase in several locations, a large late-fall/early-winter surge in 2020/2021, the rise and fall of the Delta variant in the summer and fall of 2021, and the early phase of the Omicron variant's dominance in winter 2021 (Figure 1A). Performance of the national ensemble forecasts varied over this period (Figure 1B). For some forecasts, the median predictions were close to the cases eventually reported, and most reported numbers fell within the 95% PIs. However, forecasts made at other times, such as January 2021 or December 2021, diverged widely from the reported data. At those times, the forecasts missed substantial decreases and increases, respectively, with reported cases falling within the 95% prediction interval for only 1-week ahead forecasts.

Figure 1. Weekly incident reported COVID-19 cases per 100K population, nationally (in black) and per state/territory/DC (in gray), over time in panel A. Panel B shows a subset of COVIDhub-4_week_ensemble forecasts (in green) over time, with the median predictions represented as lines and points and the 95% prediction intervals in bands. Reported incident cases (counts per week) are shown in gray. In both plots, the black, dashed vertical line shows the date that public communication of the case forecasts was paused.

Aggregate performance

We evaluated aggregate forecast performance with two metrics: Weighted Interval Score (WIS), a proper score considering both precision and accuracy, and prediction interval coverage, an indicator of forecast uncertainty. Lower WIS values reflect forecasts with probability mass closer to observed values. We assessed scaled pairwise

WIS relative to the baseline model (referred to throughout as relative WIS, or rWIS) for national and state/territory/DC forecasts (Figure 2). A rWIS less than one indicates performance that is better than the baseline model.

Figure 2: Percent of weeks with complete submissions for all sets of team forecasts, scaled, pairwise relative Weighted Interval Score (rWIS; see *Methods* for description), observed 95% prediction interval coverage, by geographical scale of submitted forecasts. Teams are sorted by increasing state/territory/DC rWIS values.

Overall, seven of 22 team's forecast models outperformed the COVIDhub-baseline model at the state/territory/DC level (i.e., had rWIS values less than 1.0), and 11 outperformed the baseline model at the national level. Six of these teams outperformed the baseline model at both scales: LNQ-ens1, COVIDhub-4_week_ensemble, USC-SI_kJalpha, LANL-GrowthRate, Microsoft-DeepSTIA, and CU-select.

PI coverage at the 95% level should be close to 95% for well calibrated forecasts. However, it was lower for most sets of team forecasts, with only one (LNQ-ens1) having coverage of at least 90% at all scales, while others were as low as 23%. PI coverage at 50% and 80% levels were also well below nominal levels for most sets of team forecasts, including the COVIDhub-4_week_ensemble (Figure 3). For the 50% prediction interval, no sets of team forecasts had coverage better than 36% at any scale. Only two sets of team forecasts had better coverage than the COVIDhub-4_week_ensemble for the geographic scales in which they submitted forecasts: LNQ-ens1 (all scales) and JHU_UNC_GAS-StatMechPool (state/territory/DC and large county levels).

Figure 3: Expected and observed coverage rates for central 50%, 80% and 95% prediction intervals aggregated over time and horizon for national forecasts (panel A), state/territory/DC forecasts (panel B), the largest county forecasts (panel C). The dashed line represents optimal expected-coverage. Team forecasts that had

closer to nominal coverage than the COVIDhub-4_week_ensemble model at all three coverage levels are labeled on the right side of the plots.

Forecast skill also showed distinct patterns across jurisdictional scales, with rWIS decreasing for larger jurisdiction scales (e.g., national vs. state/territory) or population sizes (e.g., larger counties vs. smaller counties, Figure 4) for most sets of team forecasts. In contrast to this general trend, for three sets of team forecasts, that pattern was inverted, one team had no distinct pattern, and the COVIDhub-4_week_ensemble had markedly consistent rWIS across all scales. Consistent with the aggregate findings, both LNQ-ens1 and COVIDhub-4_week_ensemble had rWIS lower than 1.0 at all scales, while LANL-GrowthRate had rWIS greater than 1.0 for smaller counties.

Figure 4: Scaled, pairwise relative Weighted Interval Score (rWIS) (see *Methods* for description) by spatial scale for sets of team forecasts that submitted forecasts for the US nation, states/territories/DC, and all US counties. WIS is averaged across all horizons. The COVIDhub-baseline model has, by definition, a rWIS of 1 (horizontal dashed line). Teams are ordered by increasing state/territory/DC rWIS with the most accurate model on the left. Points for each team are staggered horizontally to show overlapping WIS values.

Performance across jurisdictions

There was additional variability in forecast skill between jurisdictions. Only two team forecasts (LNQ-ens1 and COVIDhub-4_week_ensemble) performed as well as or better than the baseline for all included states and territories (Figure 5). Variation was higher between team forecasts than between specific jurisdictions, but the baseline model tended to outperform more models in some jurisdictions (e.g., the baseline was better in Colorado, Kansas, Puerto Rico) than in others (e.g., the baseline was worse in Mississippi, South Carolina, West Virginia).

Figure 5: Scaled, pairwise relative Weighted Interval Score (rWIS; see *Methods* for description) by location for national and state/territory/DC forecasts, averaged across all horizons through the entire analysis period. National estimates are displayed first, followed by jurisdictions in alphabetical order. Team forecasts are ordered by increasing average state/territory/DC rWIS.

Performance over time

While rWIS varied between team forecasts and jurisdictions, it varied even more over time (Figure 6). For example, all models had relatively high WIS in December 2020-January 2021 and low WIS in June 2021. Prediction interval coverage also varied between teams and over time, with most team forecasts exhibiting times of low coverage. Across most time points, the baseline model outperformed many team forecasts, including the COVIDhub-4_week_ensemble, though the ensemble more often outperformed the baseline in both metrics at the national, state/territory, and large county scales. Increased WIS and decreased prediction interval coverage generally occurred with increasing case counts, such as in the fall of 2020 and summer of 2021. The worst performance was in the early Omicron wave in the winter of 2021. For the last set of ensemble forecasts posted by CDC in December 2021 (https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasts-cases.html), the WIS reached the highest level ever for all scales and the reported case numbers were outside the 95% prediction interval for most locations at every forecast horizon.

Figure 6: Forecast accuracy over time, aggregated by geographic units, forecast horizon, and prediction date. Panels A-C show average Weighted Interval Score (WIS); panels D-F show 95% prediction interval coverage. The black, dashed vertical line in all panels shows the date that public communication of the case forecasts was paused. The black, dashed horizontal line in panels D-F shows nominal 95% interval coverage. National

level forecasts are presented in A and D, state/territory/DC forecasts in B and E and large county level forecasts in C and F.

To further investigate these temporal patterns in performance, we first classified each forecast week as *increasing*, *peak*, *decreasing*, or *nadir* based on the estimated time-varying reproduction number for that given week and jurisdiction. We then fitted Gaussian generalized estimating equations (GEE) models for each set of team forecasts, using a normalized, log transformed WIS value per forecast time and location as the model outcome. The regression models were adjusted for each prediction horizon and included a natural spline with two degrees of freedom for the time/state reported case counts to adjust for intrinsic increases in WIS due to higher values in reported cases (see S6). In agreement with the aggregated results (Figure 2), we found that the expected WIS at the mean number of case counts across all jurisdictions was lower than the baseline for the better performing models (6 team forecasts and the ensemble) and higher than the baseline for others (8 team forecasts).

Forecasts skill also varied across epidemic phases (Figure 7B). Compared to the baseline model across all phases, overall skill for most models was better in nadir and peak phases and worse in increasing and decreasing phases. LNQ-ens1 and the COVIDhub ensemble outperformed the baseline model in all epidemic phases between August 1, 2020 and January 15, 2022, while several other team models outperformed the baseline in some phases.

Figure 7. Estimated marginal mean Weighted Interval Score (WIS) and 95% confidence intervals for mean cases from team-specific GEE models for all 51 jurisdictions (Panel A). The 95% confidence intervals for the COVIDhub-baseline model are shown in dashed red vertical lines. Panel B presents each team's estimated marginal mean WIS per phase, scaled to the COVIDhub-baseline model's estimated marginal mean WIS for all epidemic phases. Teams with higher estimated marginal mean WIS values (i.e., greater than 1.0) are

presented in shades of orange while teams with lower estimated marginal mean WIS (i.e., less than 1.0) are shown in shades of green. Forecasts for a team in a particular phase are marked with an asterisk (*) if the 80% confidence interval of the expected WIS outcome (normalized and on the log scale) was estimated by a model to be lower than the expected WIS of the COVIDhub-baseline model for all phases.

To examine whether our results were affected by reporting anomalies, we also conducted sensitivity analyses for data revisions, when data were revised at a later date, and for outlier data points, when reported cases were outside of weekly expected ranges (see S2). We first identified weeks in which revised case counts as of April 2, 2022 differed from the case counts initially reported for that week, excluded them from the dataset, and reran the GEE models. With this partial dataset, the results were essentially unchanged. Next, we identified outliers as reported case counts outside of the expected range by at least two of the three following algorithms: a rolling median, a seasonal trend decomposition, and a seasonal trend decomposition without a seasonality term, each method over a 21-day window. Approximately 3% of weeks (686 of 27,489 total week-location combinations in the analysis period) had at least one day of reported cases identified as an outlier. We then excluded the weeks with outliers and the week following an outlier and reran the GEE models. This sensitivity analysis had comparable results to the models with the full data (see S2 Figure 2.3, Panel A.).

Discussion

We evaluated performance of 9.7 million COVID-19 case forecasts at multiple geospatial scales in the US over approximately a year and a half. Real-time analyses and those presented here revealed important limitations in these forecasts. Forecast prediction intervals were largely over-confident, that is, prediction interval coverage was lower than the nominal value, particularly when case numbers were changing rapidly and forecasts could have been most useful. Few team forecasts outperformed a relatively simple and minimally informative baseline model. Forecast skill degraded for smaller geographic scales where forecasts could potentially be most useful.

Forecast skill was also lowest when case counts were changing the most, in phases of increasing or decreasing transmission. These limitations of case forecasts indicate key areas for improvement and important reasons to use case forecasts with caution.

Several technical challenges for forecasts were evident in these analyses. First, cases are a relatively early indicator of transmission, with no clear leading signal in traditional public health surveillance data (e.g., unlike for death forecasts, where case counts themselves can provide information for predicting future deaths). While non-traditional data sources may provide a useful predecessor to changing population case counts, the evidence from previous work is unclear. For example, internet searches, medical claims, and online surveys have been used to modestly improve case forecast accuracy relative to models without those data (19). Estimating case counts using both wastewater and clinical surveillance data has shown mixed results (20–23). Additional integration of temporal dynamics could also be helpful. The case forecasts analyzed here were developed and evaluated based on the date when cases were reported, not when individuals were infected, became ill, sought care, or were tested. Additional detail on those dates could enable models to better capture the current dynamics using nowcasting approaches giving earlier signals of change.

Second, and likely related to the challenge of cases being an early indicator, the models had substantial variation in skill between epidemic phases. In general, forecast skill was worst for the increasing phase followed by the decreasing phase. In many of these periods of low performance (e.g., the 2020-2021 winter, Delta, and Omicron waves), the COVIDhub ensemble predicted possible or probable increases or decreases, but not at the rate that actually occurred. This effect may be even stronger than our results show as they rely on a comparison to the baseline which, by definition, does not predict change. While epidemic phase is unknown in real time, it too can be estimated, and these results and others suggest that accounting for epidemic phase when making predictions could improve the forecast skill of ensemble models (24,25). Additional data, as discussed above, or model components associated with distinct phases could also help improve predictive capabilities. Seasonal changes in transmission biology and human behavior, emergence of variants, and changing mitigation behavior all contribute to transmission dynamics. While some forecasting models incorporate seasonality and variants, integration of human behavior to characterize the link between behavior and transmission has lagged (13,26–28). Ensemble approaches offer another opportunity to mitigate phase-specific differences. Team modeling skill across phases was highly heterogeneous, but two ensemble approaches were better than the baseline in all phases.

Another challenge across most forecasts was overconfidence, a pattern seen with other infectious disease forecasts (4,18). The baseline model predicted a flat trend, yet it outperformed many sets of team forecasts in the increasing and decreasing phase only because its predictions had high uncertainty around that flat trend. The COVIDhub ensemble performance, on the other hand, benefitted by combining uncertainty across multiple models, yet, like the constituent models, also exhibited overconfidence. The temporal and phase-specific analyses suggest that it is, during rapid increases and decreases, that model overconfidence is most pronounced. Previous infectious disease forecasting work has shown that ensembles tend to have wider prediction intervals that are more likely to capture the eventually reported outcome and thus reduce overconfidence compared to their constituent models (4,18). Wider prediction intervals, reflecting increased uncertainty, can mediate some impacts of overconfidence. However, forecasts would be most useful if they were both reliable and informative - that is, if they could accurately capture the uncertainty, while also providing more precise estimates, rather than merely increased uncertainty (29,30).

Finally, while forecasts would be most actionable at local scales, performance was generally worse for smaller than larger jurisdictions. Other infectious disease forecasting systems have found better forecast skill at smaller geographic scales, likely because local transmission dynamics (e.g., a county) are a better predictor of local than aggregate transmission (e.g., a state) (31). We compared WIS across scales by comparison to the baseline model to adjust for missing forecasts and for WIS scaling relative to the magnitude of observed outcomes. After those adjustments, population size had a clear association with forecast , likely reflecting the relative role of stochastic dynamics. For better local forecasts, models may need to explicitly account for stochasticity. Forecasts could also be improved by better leveraging spatial information, such as dynamics in neighboring counties or nearest urban centers. Local forecasts remain a key public health need, as local forecasts are more likely to reflect local conditions and motivate local mitigation action.

Overall, these findings, as well as the real-time evaluations, indicated that COVID-19 case forecasts were not reliable as a single indicator for pandemic response of a novel pathogen. Similar to other forecasting studies, we found that the ensemble was among the most reliable forecasts (3,4,18,32), outperformed only by LNQ-ens1 across the metrics evaluated here. Thus, while the overall best forecasts had poor performance at key times, other forecasts were often even worse at these same time points. Weighted (or trained) ensembles offer another potential avenue for improvement (33–35), but the version implemented here did not outperform the simple, median ensemble, likely reflecting limited historical data (36) and variation in team forecast submissions (37,38).

While COVID-19 deaths are a more lagging indicator of infections than case reports, and so may be less useful as an input to public health decision making, forecasts of deaths have generally been more reliable (18). Similarly, COVID-19 hospitalization forecasts in France have also shown high forecast skill (39). Better performing US death and French hospitalization forecasts share one factor in common: models generally used local case reports as an input to inform their forecasts. While public health decision making should not rely on case forecasts alone, they may still be helpful in the context of other important indicators, such as the case, hospitalizations, and death reports. Nowcasts of reports and real-time estimates of the effective reproductive number can also provide insight on current dynamics (40–43). Together, a suite of indicators is more informative for outbreak response than a leading indicator alone.

The analysis presented here includes important findings about real-time applied forecasting in an emerging pandemic to inform pandemic response rather than to address specific research aims of improving predictions. Several factors limit the strength of our findings and ability to understand underlying mechanisms of predictive performance. Notably, we compared the forecasts to a changing record of reported cases. Throughout the COVID-19 outbreak, cases have been reported with jurisdiction- and time-varying delays and have been revised over time, resulting in varying forecast targets. In addition, the definition of a reported COVID-19 case also changed over time and varied between states. These changes were a result of many factors, including laboratory capacity and implementation of home-based testing, and may have affected forecast skill in other ways. Our sensitivity analyses found no qualitative differences in our main findings when we excluded forecasts for time points with revised data or when we excluded outlier data points. Nevertheless, forecasting teams were greatly impacted by the evolving landscape of COVID-19 case surveillance. More timely and consistent reports likely would improve both the process of making forecasts and forecast skill.

The overall goal of the COVID-19 Forecast Hub was to provide forecasts in near real-time for decision making. While the collaborative efforts of the Hub achieved this goal despite a changing epidemic landscape, nevertheless, the e open nature of COVID-19 forecasting also limits understanding the drivers of forecast performance. Many teams participated at different times, some intermittently, and provided varied and limited descriptions of their forecast methods. While we were able to adjust our evaluation for differences in in varying submissions, we are unable to assess the underlying impact of modeling approaches on performance since we do not have the granular details on forecast methods and how they evolved over time for all team forecasts. For example, the LNQ-ens1, which outperformed all other forecasts by most metrics, only submitted forecasts for approximately two thirds of the analysis period and stopped in June 2021 (prior to the Delta wave). The model is described as a combination of three machine learning models, leveraging other embedded models and datasets, with weights that "are chosen by hand each week based on performance in the previous week" (see LNQ-ens1 metadata, https://github.com/reichlab/covid19-forecast-hub/blob/b12f916abc859bf59ea584b64f53afc2982042fd/data-processed/LNQ-ens1/metadata-LNQ-ens1.txt, at (44)). The ensemble approach used in the LNQ-ens1 model building likely contributed to the overall performance. However, several other ensemble models had lower performance than the LNQ-ens1 model; we are unable to assess whether LNQ-ens1 performance gains were due to a particular component model or dataset, the hand weighting procedure, or something else. The brief descriptions submitted to the COVID-19 Forecast Hub, such as for the LNQ-ens1, must include a summary of the methods used and may indicate a variety of unique features such as input data, parameters, model fitting, etc. (44). However, the level of detail provided in these descriptions varies between teams, and we do not have enough information to determine which aspects of individual models were important determinants of forecast performance. To elucidate associations between modeling approaches and forecast skill, additional research is needed. Future work to support improved forecasting will require assessing the impact of specific features (e.g., through ablation analyses) using retrospective, sable data systems and retrospective evaluation of the full forecasting process (e.g., from data wrangling to final forecast production).

Infectious disease forecasting continues to present many challenges and opportunities for improving outbreak response. Forecasts should be leading indicators of future activity and, while the COVID-19 case ensemble forecasts were good leading indicators at many points in time; they were unreliable, especially during periods of rapid change. Case data were integrated in COVID-19 mortality forecasts, which proved to be more reliable, likely in part due to reported cases being leading indicators of reported deaths (18,45). However, because deaths are a lagging indicator, death forecasts are less useful for short-term outbreak responses. Evaluation of the case forecasts provided insight on limitations of early forecasts and research avenues for improving them. These insights and the real-time forecasts provided by this effort were the product of large-scale collaboration between researchers and public health responders to confront the COVID-19 pandemic. Learning from and improving forecasting for COVID-19, other infectious diseases, and future pandemics will benefit from continuing and expanding these collaborative efforts.

Methods

The US COVID-19 Forecast Hub (46) is a consortium of researchers that develop and share forecasts of COVID-19 reported cases, hospitalizations, and deaths with the goal of leveraging information from individual models that predict the near-term burden of COVID-19 in the United States. Teams that submitted models to the US COVID-19 Forecast Hub used a wide variety of methodology and data (S1, Table S1). Beyond serving as a repository for forecasts, submitted data were also aggregated by scientists at the COVID-19 Forecast Hub to generate two models that we included in this analysis: the COVIDhub-4_week_ensemble and the COVIDhub-trained_ensemble. Since the beginning of the COVID-19 Forecast Hub, the quantile predictions from each week's submitted models were used as input data for the COVIDhub-4_week_ensemble. Ensemble aggregation methods evolved over time; for this analysis period, the ensemble forecast was calculated as the median across forecasts from all models at each quantile level. Additionally, beginning on February 1, 2021, the COVID-19 Forecast Hub also generated a weighted ensemble (COVIDhub-trained_ensemble). Models were selected for weighted ensemble inclusion based on their past performance over various window period and given a weight prior to aggregation. The methodology evolved over time and details are available on the model's metadata file on the COVID-19 Forecast Hub GitHub repository (see *Data and code availability and reporting guidelines*).

The COVID-19 Forecast Hub, and death forecasts submitted to the Hub have been described in detail elsewhere (8,17,18). The Hub's incident COVID-19 case forecasts, which were first solicited in July 2020, have similar submission requirements to the death forecasts. Important differences include an expanded geographical scale (national; state, territory, and DC; and county levels) and reduced number of required quantiles in the probability distribution (7 quantiles in total: 0.025, 0.10, 0.25, 0.50, 0.75, 0.90, and 0.975). Predictions for weekly incident COVID-19 cases can be submitted for up to 8 weeks in the future, although our analysis only includes predictions made for 1-4 weeks into the future.

We evaluated submitted forecasts between July 28, 2020, and December 21, 2021 (2020 epi week [EW] 31 – 2021 EW 51), which encompasses 73 weeks. Because forecasts were submitted at multiple geographic scales, we conducted separate analyses for 1) national forecasts, 2) state, territory, and DC forecasts, 3) county level forecasts, and 4) sets of team forecasts for all three geographic scales. When appropriate, we compared forecast performance to that of a naïve model, created by the COVID-19 Forecast Hub, the COVIDhub-baseline. The COVIDhub-baseline model, created each week, was designed to be a neutral model to provide a simple reference point of comparison for all models. This baseline model forecasts a predictive median incidence equal to the number of reported cases in the most recent week, with uncertainty based on the empirical distribution of previous differences between the median and observed values (18).

Inclusion criteria

Teams were included in the evaluation when they submitted forecasts with a complete set of quantiles for each 1- through 4-week ahead target predictions. Additionally, teams must have met the following inclusion criteria:

1. had predictions for at least 50 locations (states, territories, or DC) for the state, territory, and DC level analyses; and for at least 75% of counties included in each population size quantile per submission week for the county-level analyses;
2. had submissions for at least 50% of the weeks included in the analysis period per location forecasted.

Teams meeting these inclusion criteria, and their submissions over time, are depicted in S1, Figure S1.

Ground Truth

Forecasts were evaluated against the reported COVID-19 case reports collated by the Johns Hopkins Center for Systems Science and Engineering (CSSE) (47). To calculate weekly incident reported cases, we subtracted the cumulative count for each Saturday from the cumulative count for the next Saturday, such that each incident weekly count reflects the number of cases reported from Sunday through Saturday in a given week. We aggregated reported counts from smaller geographic units into their larger unit. For example, counts in a given state are the aggregate of the county level reported counts and national counts are the sum of all states, territories, and DC.

CSSE reports data in real-time. Thus, data may be revised if the reporting health system makes public updates to their surveillance data. At times, such revisions may result in negative daily counts or in increases to case counts if the date of cases is shifted from one day to another or the definition of a reportable case is changed. We examined the percent change between the first reported cases in each state, DC, and territory per date relative to the counts in the surveillance file from April 2, 2022. We also assessed the influence of revised data on the final model outcomes (see S2) and the presence of negative case counts in the timeseries. Less than 1 percent of time points in the analysis period had negative daily case counts in the largest US counties. Negative counts were observed at the state/territory level only twice: in Missouri during the week of April 17, 2021, and Virgin Islands during the week ending October 10, 2020. The state of Florida reported 0 cases on November 27, 2021. We excluded all weeks and locations with negative counts as well as the week with 0 incidence in Florida in our primary analyses.

Additionally, we also examined whether a reported case count was an outlier in the case trend for each state. Anomalies in case data trends have not been uncommon throughout the pandemic, as reporting entities have uploaded large batches of surveillance data on a single day. To assess whether cases were outside of the expected range of reported cases over time, we applied three outlier detection algorithms, each with a 21-day window: a rolling median, a seasonal trend decomposition, and a seasonal trend decomposition without a seasonality term. We then classified a given count as an outlier if it was detected as such by at least two of the three algorithms. Using these data, we ran several sensitivity analyses to assess the likely impact of anomalous data points on model performance. Sensitivity analyses examining the robustness of our findings to reporting anomalies are presented in S2.

Additional information about the CSSE data, and revisions to the dataset, is publicly available on a GitHub repository:
https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data.

Forecast locations

Forecasts for incident cases were submitted for the national level, 50 states, 5 territories (American Samoa, Guam, the Northern Mariana Islands, Puerto Rico, and the US Virgin Islands), the DC, and 3,142 US counties. We excluded two counties in Alaska because they were not forecasted by most sets of team forecasts (Federal Information Processing Standard code 02063 and 02066 were excluded). Because fewer teams submitted forecasts for American Samoa, Guam, the Northern Mariana Islands, we excluded these territories from the analysis. Some teams treated DC as both a county and a jurisdiction, so we excluded DC from the county forecasts. In addition, because county population size and transmission are correlated and case counts and forecast performance are also correlated, we grouped counties into 5 quantiles based on their population sizes, with cut points at 8,908; 18,662; 36,742; and 93,230 people; most analyses used forecasts from the quantile with the largest population size (n=628). We hypothesized that small counties would be more likely to have better forecast accuracy because they had zero or very few reported cases. We thus chose to stratify counties by size to minimize any bias from aggregation. Performance results for most county forecasts are presented in S3.

Defining epidemic phases

For every state and DC, we independently classified each forecast week based on the estimated time-varying reproduction number (R_t) for that given week. State-level R_t estimates were obtained from https://github.com/epiforecasts/epiforecasts.github.io (48). We extracted the R_t estimate for the Wednesday of each week from all available files. Because R_t estimates were updated on a rolling basis in near real time, there were multiple estimates generated for the same date; we calculated the median estimated R_t per date for the upper and lower 90% credible interval and the median value (August 1, 2020 – January 15, 2022, or 2020 EW 31 – 2022 EW 2, reflecting 77 weeks in total). Each forecast week was then classified into one of the following categories based on the R_t estimates: *increasing*, *peak*, *decreasing*, *nadir.*

Increasing and *decreasing* phases reflect weeks in which R_t had a 90% probability of being greater than or less than 1.0, respectively. There were several periods of rapid transmission in certain jurisdictions where R_t dipped above/below the 1.0 threshold but did not remain on an upward or downward trajectory. Thus, we classified weeks between two increasing phases as *increasing* and weeks between two decreasing periods as *decreasing*. Weeks between increasing and decreasing phases were classified as *peaks*, whereas *nadirs* were defined as periods between decreasing and increasing phases. Periods at the beginning or the end of an analysis period were classified as a continuation of whichever phase preceded or followed them. Graphical depictions of R_t are provided in S4 and show general concordance between R_t and reported cases.

Evaluation methodology

We evaluated probabilistic forecast accuracy using two different metrics, empirical prediction interval coverage rates and weighted interval scores (WIS) (49). Coverage was calculated by determining the frequency with which the prediction interval contained the eventually observed outcome for the 50%, 80% and 95% intervals. WIS reflects a weighted estimate of sharpness (i.e., the range of the predicted interval) and calibration (i.e., precision or error) across the three prediction intervals and the median prediction, with higher WIS and indicating lower forecast skill. Importantly, WIS is highly correlated with the magnitude of observed and forecasted values. We used mean absolute WIS to assess forecast accuracy over time and mean relative WIS (rWIS) to access forecast accuracy over space. Relative WIS was estimated by calculating the geometric mean of WIS across all sets of team forecasts and scaling that value to the WIS of a naïve model, the COVID-hub baseline. This approach eases interpretation, where values greater than 1.0 reflected worse accuracy than the baseline model and values below 1.0 reflected better model performance. Additionally, the pairwise relative comparison helps account for missing forecasts. Both coverage and WIS have been described in detail elsewhere (18,49). Horizon specific results for national, state/territory/DC, and large counties are presented in S5.

To assess the association between WIS and epidemic phase for each team, we fitted separate Gaussian generalized estimating equation (GEE) models per team (equation 1) with an independent working correlation structure at the state-level. This structure assumes that observations are not correlated over time in a state (denoted as l in the equations below). Cases and weighted interval scores were log transformed and then standardized (subtracting the mean and dividing by the standard deviation) prior to fitting the model, as this transformation yielded more computationally and numerically stable estimates. We define those resulting variables as stdWIS and stdCases. The expected value for a standardized WIS for time (*t)* and location (*l)*, with forecasts from a given team's model, is as follows:

$$log\left(stdWIS_{t,l,h}\right) = \beta_0 + \alpha_{p[t,l]} + \gamma_h + ns\left(log\left(stdCases_{t,l,h}\right)\right) + \epsilon_{t,l,h} \quad \textbf{(1)}$$

Where $p[t,l]$ is an index that reflects the phase of each time (t) and location (l), (h) is the horizon of the forecast in weeks, and $ns(\cdot)$ represents a natural spline with two degrees of freedom. Using a regression model allows us to summarize patterns of overall average performance between teams while accounting for high correlation and variation in the scores. Comparisons of rWIS, in contrast, do not allow for formal inference on the differences in performance between teams. Prior to applying this regression model structure, our model building approach included exploratory analysis of several structures appropriate for longitudinal analysis. We examined model residuals, influential observations, goodness of fit metrics, and the impact of changing the functional form of the variables included in the model.

The inclusion of reported cases in models permitted flexible adjustment for the wide range in cases between and within jurisdictions, which led to a wide range of possible WIS values, as WIS values tend to be higher when counts are higher. Expected WIS values were computed by first obtaining a marginal mean from the GEE model and then undoing the transformations by exponentiating and un-standardizing the marginal mean. This was done separately for each team for all phases and for each team and each phase individually (see S6 for estimated team-specific marginal mean WIS relative to reported case counts). Additionally, we calculated whether the 80% confidence interval (based on Gaussian distributional assumptions) for each team's expected WIS outcome (on the log-scale and normalized, as described above) was less than the baseline model for all phases (i.e., the marginal mean WIS for the baseline model).

Data and code availability and reporting guidelines

The forecasts from models used in this paper are available from the COVID-19 Forecast Hub GitHub repository (https://github.com/reichlab/covid19-forecast-hub) (8) and the Zoltar forecast archive (https://zoltardata.com/project/44) (50). The code used to generate all figures and tables in the manuscript is available in a public repository (https://github.com/cdcepi/Evaluation-of-case-forecasts-submitted-to-COVID19-Forecast-Hub). All analyses were conducted using the R language for statistical computing (v 4.0.3) (51), and the following packages were used for the main analyses: *scoringutils* (52), *covidhubUtils* (53), *geepack* (54). Additionally, we included the EPIFORGE 2020 reporting guideline checklist in S7 to indicate each page in this evaluation that corresponds to each specific recommendation (15).

This activity was reviewed by CDC and was conducted consistent with applicable federal law and CDC policy.

CDC disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

References

1. Biggerstaff M, Johansson M, Alper D, Brooks LC, Chakraborty P, Farrow DC, et al. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. Epidemics. 2018 Sep 1;24:26–33.
2. McGowan CJ, Biggerstaff M, Johansson M, Apfeldorf KM, Ben-Nun M, Brooks L, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. Sci Rep 2019 91. 2019 Jan 24;9(1):1–13.
3. Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proc Natl Acad Sci U S A. 2019 Feb 19;116(8):3146–54.
4. Johansson MA, Apfeldorf KM, Dobson S, Devita J, Buczak AL, Baugher B, et al. An open challenge to advance probabilistic forecasting for dengue epidemics. Proc Natl Acad Sci. 2019 Nov 26;116(48):24268–74.
5. Holcomb KM, Barker CM, Keyel Wadsworth Center Matteo Marcantonio AC, Childs ML, Gorris ME, Hamins-Puértolas M, et al. Evaluation of an open forecasting challenge to assess skill of West Nile virus neuroinvasive disease prediction. 2022 Aug 26 [cited 2022 Dec 16]; Available from: https://www.researchsquare.com
6. Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, et al. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. Epidemics. 2018 Mar 1;22:13–21.
7. Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, et al. Applying infectious disease forecasting to public health: A path forward using influenza forecasting examples. BMC Public Health. 2019 Dec 10;19(1):1–12.
8. Cramer EY, Huang Y, Wang Y, Ray EL, Cornell M, Bracher J, et al. The United States COVID-19 Forecast Hub dataset. Sci Data. 2022 Aug 1;9(1):462.

9. Centers for Disease Control and Prevention. COVID-19 Forecasts: Cases | CDC [Internet]. [cited 2022 Dec 16]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasts-cases.html
10. Biggerstaff M, Slayton RB, Johansson MA, Butler JC. Improving Pandemic Response: Employing Mathematical Modeling to Confront Coronavirus Disease 2019. Clin Infect Dis. 2022 Mar 9;74(5):913–7.
11. Bertsimas D, Boussioux L, Cory-Wright R, Delarue A, Digalakis V, Jacquillat A, et al. From predictions to prescriptions: A data-driven response to COVID-19. Health Care Manag Sci. 2021 Jun 1;24(2):253–72.
12. Padilla L, Hosseinpour H, Fygenson R, Howell J, Chunara R, Bertini E. Impact of COVID-19 forecast visualizations on pandemic risk perceptions. Sci Rep 2022 121. 2022 Feb 7;12(1):1–14.
13. Eksin C, Paarporn K, Weitz JS. Systematic biases in disease forecasting – The role of behavior change. Epidemics. 2019 Jun 1;27:96–105.
14. Reich NG, Tibshirani RJ, Ray EL, Rosenfeld R. On the predictability of COVID-19 - International Institute of Forecasters [Internet]. [cited 2022 Dec 16]. Available from: https://forecasters.org/blog/2021/09/28/on-the-predictability-of-covid-19/
15. Pollett S, Johansson MA, Reich NG, Brett-Major D, Del Valle SY, Venkatramanan S, et al. Recommended reporting items for epidemic forecasting and prediction research: The EPIFORGE 2020 guidelines. PLOS Med. 2021 Oct 1;18(10):e1003793.
16. Nixon K, Jindal S, Parker F, Reich NG, Ghobadi K, Lee EC, et al. An evaluation of prospective COVID-19 modelling studies in the USA: from data to science translation. Lancet Digit Health. 2022 Oct 1;4(10):e738–47.
17. Ray EL, Wattanachit N, Niemi J, Kanji AH, House K, Cramer EY, et al. Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. medRxiv. 2020 Aug 22;2020.08.19.20177493.
18. Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proc Natl Acad Sci. 2022 Apr 12;119(15):e2113561119.
19. McDonald DJ, Bien J, Green A, Hu AJ, DeFries N, Hyun S, et al. Can auxiliary indicators improve COVID-19

forecasting and hotspot prediction? Proc Natl Acad Sci. 2021 Dec 21;118(51):e2111453118.

20. McMahan CS, Self S, Rennert L, Kalbaugh C, Kriebel D, Graves D, et al. COVID-19 wastewater epidemiology: a model to estimate infected populations. Lancet Planet Health. 2021 Dec 1;5(12):e874–81.
21. Nourbakhsh S, Fazil A, Li M, Mangat CS, Peterson SW, Daigle J, et al. A wastewater-based epidemic model for SARS-CoV-2 with application to three Canadian cities. Epidemics. 2022 Jun 1;39:100560.
22. Proverbio D, Kemp F, Magni S, Ogorzaly L, Cauchie HM, Gonçalves J, et al. Model-based assessment of COVID-19 epidemic dynamics by wastewater analysis. Sci Total Environ. 2022 Jun 25;827:154235.
23. Cao Y, Francis R. On forecasting the community-level COVID-19 cases from the concentration of SARS-CoV-2 in wastewater. Sci Total Environ. 2021 Sep 10;786:147451.
24. Adiga A, Kaur G, Wang L, Hurt B, Porebski P, Venkatramanan S, et al. Phase-Informed Bayesian Ensemble Models Improve Performance of Covid-19 Forecasts. In Thirty-Fifth Annual Conference on Innovative Applications of Artificial Intelligence; 2023.
25. Adiga A, Kaur G, Hurt B, Wang L, Porebski P, Venkatramanan S, et al. Enhancing COVID-19 Ensemble Forecasting Model Performance Using Auxiliary Data Sources. In IEEE International Conference on Big Data; 2022.
26. Funk S, Salathé M, Jansen V a a. Modelling the influence of human behaviour on the spread of infectious diseases: a review. J R Soc Interface R Soc. 2010;7(50):1247–56.
27. Moran KR, Fairchild G, Generous N, Hickmann K, Osthus D, Priedhorsky R, et al. Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast. J Infect Dis. 2016 Dec 1;214(suppl_4):S404–8.
28. Buckee C, Noor A, Sattenspiel L. Thinking clearly about social aspects of infectious disease transmission. Nat 2021 5957866. 2021 Jun 30;595(7866):205–13.
29. Bodner K, Fortin MJ, Molnár PK. Making predictive modelling ART: accurate, reliable, and transparent. Ecosphere. 2020;11(6):e03160.
30. Yanvi I, Foster DP. Graininess of Judgment Under Uncertainty: An Accuracy–Informativeness Trade-Off. J Exp

Psychol Gen. 1995 Dec;124(4):24–432.

31. Reis J, Yamana T, Kandula S, Shaman J. Superensemble forecast of respiratory syncytial virus outbreaks at national, regional, and state levels in the United States. Epidemics. 2019 Mar 1;26:1–8.
32. Oidtman RJ, Omodei E, Kraemer MUG, Castañeda-Orjuela CA, Cruz-Rivera E, Misnaza-Castrillón S, et al. Trade-offs between individual and ensemble forecasts of an emerging infectious disease. Nat Commun. 2021 Sep 10;12(1):5379.
33. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. PLOS Comput Biol. 2018 Jun 1;14(6):e1006134.
34. Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density ensembles. PLOS Comput Biol. 2018 Feb 1;14(2):e1005910.
35. McAndrew T, Reich NG. Adaptively stacking ensembles for influenza forecasting. Stat Med. 2021 Dec 30;40(30):6931–52.
36. Smith J, Wallis KF. A simple explanation of the forecast combination puzzle. Oxf Bull Econ Stat. 2009;71(3).
37. Ray EL, Brooks LC, Bien J, Biggerstaff M, Bosse NI, Bracher J, et al. Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. Int J Forecast. 2022 Jul 1;
38. Taylor JW, Taylor KS. Combining probabilistic forecasts of COVID-19 mortality in the United States. Eur J Oper Res. 2023;304(1).
39. Paireau J, Andronico A, Hozé N, Layan M, Crépey P, Roumagnac A, et al. An ensemble model based on early predictors to forecast COVID-19 health care demand in France. Proc Natl Acad Sci U S A [Internet]. 2022 May 3 [cited 2022 Dec 16];119(18). Available from: https://doi.org/10.1073/pnas.2103302119
40. Morozova O, Li ZR, Crawford FW. One year of modeling and forecasting COVID-19 transmission to support policymakers in Connecticut. Sci Rep. 123AD;11:20271.
41. Greene SK, McGough SF, Culp GM, Graf LE, Lipsitch M, Menzies NA, et al. Nowcasting for Real-Time COVID-19 Tracking in New York City: An Evaluation Using Reportable Disease Data From Early in the Pandemic.

JMIR Public Health Surveill 202171e25538. 2021 Jan 15;7(1):e25538.

42. covidestim: COVID-19 nowcasting [Internet]. [cited 2023 Jan 9]. Available from: https://covidestim.org/
43. Abbott S, Sherratt K, Bevan J, Gibbs H, Hellewell J, Munday J, et al. Temporal variation in transmission during the COVID-19 outbreak [Internet]. Vol. 2020 [cited 2023 Jan 9]. Available from: https://epiforecasts.io/covid/
44. reichlab/covid19-forecast-hub [Internet]. The Reich Lab at UMass-Amherst; 2022 [cited 2023 Jan 10]. Available from: https://github.com/reichlab/covid19-forecast-hub
45. Bracher J, Wolffram D, Deuschel J, Görgen K, Ketterer JL, Ullrich A, et al. National and subnational short-term forecasting of COVID-19 in Germany and Poland during early 2021. Commun Med 2022 21. 2022 Oct 31;2(1):1–17.
46. COVID 19 forecast hub [Internet]. [cited 2022 Dec 16]. Available from: https://covid19forecasthub.org/
47. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020 May 1;20(5):533–4.
48. Abbott S, Bennett C, Hickson J, Allen J, Sherratt K, Funk S. National and Subnational Estimates of the Covid 19 Reproduction Number (R) for the United States of America Based on Test Results. Harvard Dataverse, V292. 2020.
49. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. PLoS Comput Biol. 2021 Feb 1;17(2).
50. Reich NG, Cornell M, Ray EL, House K, Le K. The Zoltar forecast archive, a tool to standardize and store interdisciplinary prediction research. Sci Data. 2021 Feb 11;8(1):59.
51. R Core Team. R: A language and environment for statistical computing. [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: https://www.R-project.org/.
52. Bosse NI, Gruson H, Cori A, van Leeuwen E, Funk S, Abbott S. Evaluating Forecasts with scoringutils in R. 2022 May 14 [cited 2022 Dec 16]; Available from: https://arxiv.org/abs/2205.07090v1
53. Wang S, Ray EL, Reich NG, Shah A. Tools for working with COVID-19 Forecast Hub data: a brief tour of the

`covidHubUtils` R package [Internet]. [cited 2022 Dec 16]. Available from: http://reichlab.io/covidHubUtils/articles/covidHubUtils-overview.html

54. Højsgaard S, Halekoh U, Yan J. The R Package geepack for Generalized Estimating Equations. J Stat Softw. 2006;15:1–11.
55. Linas BP, Xiao J, Dalgic OO, Mueller PP, Adee M, Aaron A, et al. Projecting COVID-19 Mortality as States Relax Nonpharmacologic Interventions. JAMA Health Forum. 2022 Apr 1;3(4):e220760.
56. Sen P, Yamana TK, Kandula S, Galanti M, Shaman J. Burden and characteristics of COVID-19 in the United States during 2020. Nature. 2021;598(7880).
57. Suchoski B, Stage S, Gurung H, Baccam P. GPU Accelerated Parallel Processing for Large-Scale Monte Carlo Analysis: COVID-19 Parameter Estimation and New Case Forecasting. Front Appl Math Stat. 2022;8.
58. Wang Y, Kim M, Yu S, Li X, Wang G, Wang L. Nonparametric estimation and inference for spatiotemporal epidemic models. J Nonparametric Stat. 2022;34(3).
59. Lemaitre JC, Grantz KH, Kaminsky J, Meredith HR, Truelove SA, Lauer SA, et al. A scenario modeling pipeline for COVID-19 emergency planning. Sci Rep 2021 111. 2021 Apr 6;11(1):1–13.
60. Karlen D. Characterizing the spread of CoViD-19. 2020 Jul 14 [cited 2023 Mar 10]; Available from: https://arxiv.org/abs/2007.07156v1
61. Castro L, Fairchild G, Michaud I, Osthus D. COFFEE: COVID-19 Forecasts using Fast Evaluations and Estimation. 2021 Oct 4 [cited 2023 Mar 10]; Available from: https://arxiv.org/abs/2110.01546v1
62. Zheng S, Gao Z, Cao W, Bian J, Liu TY. HierST: A Unified Hierarchical Spatialoral Framework for COVID-19 Trend Forecasting. Int Conf Inf Knowl Manag Proc. 2021 Oct 26;4383–92.
63. Davis JT, Chinazzi M, Perra N, Mu K, Pastore y Piontti A, Ajelli M, et al. Cryptic transmission of SARS-CoV-2 and the first COVID-19 wave. Nat 2021 6007887. 2021 Oct 25;600(7887):127–32.
64. Srivastava A. The Variations of SIkJalpha Model for COVID-19 Forecasting and Scenario Projections. 2022 Jul 6 [cited 2023 Mar 10]; Available from: https://arxiv.org/abs/2207.02919v1
65. Adiga A, Wang L, Hurt B, Peddireddy A, Porebski P, Venkatramanan S, et al. All Models Are Useful: Bayesian

Ensembling for Robust High Resolution COVID-19 Forecasting. Proc ACM SIGKDD Int Conf Knowl Discov Data Min. 2021 Aug 14;2505–13.

Supporting information captions

Supporting Information 1: Team submissions, methods, and data

SI Figure 1.1. Forecasts submitted over time at the national, state-territory-DC level in panel A and at the country scale in Panel B. The number of forecasted locations submitted each week nationally or at the state, territory and DC level is included, while the country level forecast submissions shows the percent of counties per quantile that were submitted each week. Sets of team forecasts meeting the inclusion criteria for this main analysis are labeled with an asterisk (*).

S1 Table 1.1. List of models evaluated, including sources for case, hospitalization, death, demographic, and mobility data when used as inputs for the given model. We evaluated 26 models contributed by 24 teams. The COVIDhub team submitted three models including the baseline model and the ensemble model. A brief description is included for each model, with a reference where available. The last column indicates whether the model made assumptions about how and whether social distancing measures were assumed to change during the period for which forecasts were made.

Supporting Information 2: Revision and outlier sensitivity analyses

S2 Figure 2.1. To assess the influence of data revisions on our evaluation of forecast skill, we compared daily differences in cumulative reported cases during the week they were first reported to reported case counts for

the same week in the complete data as of April 2, 2022. In total 721 weeks had at least one day with a revised case count (17% of all weeks, n=4,241 weeks) and revisions occurred in 43 of 51 jurisdictions. These jurisdiction specific plots compare cases reported as of the date in the subtitle (in red) compared to cases reported as of April 2, 2022 (in black).

S2 Figure 2.2. After identifying weeks with revised case counts, we then excluded them from the dataset and reran the GEE models and estimated the marginal mean Weighted Interval Score (WIS). Panel A shows the estimated marginal mean WIS and 95% confidence intervals for mean cases from team-specific GEE models for all 48 jurisdictions from this sensitivity analysis. The 95% confidence intervals for the COVIDhub-baseline model are shown in dashed red vertical lines. Panel B presents each team's estimated marginal mean WIS per phase, scaled to the COVIDhub-baseline model's estimated marginal mean WIS for all epidemic phases, using the dataset with excluded week. Teams with higher estimated marginal mean WIS values (i.e., greater than 1.0) are presented in shades of orange while teams with lower estimated marginal mean WIS (i.e., less than 1.0) are shown in shades of green. Team forecasts are denoted with an asterisk (*) if the 80% confidence interval of the expected WIS outcome (normalized and on the log scale) was estimated by a model to be lower than the expected WIS of the COVIDhub-baseline model for all phases.

S2 Figure 2.3. Outliers were defined as non-revised reported case counts that were outside of the expected range by at least two of the three algorithms: a rolling median, a seasonal trend decomposition, and a seasonal trend decomposition without a seasonality term. Each method used a 21-day window. Approximately three percent of weeks (686 of 27,489 total weeks in the analysis period) had at least one day of reported cases identified as an outlier.

Supporting Information 3: Incident COVID-19 case forecasts were submitted for all US counties. The plots shown here depicted average, scaled pairwise Weighted Interval Score (WIS; see *Methods* for description), 95% coverage, and submissions (S3 Figure 3.1), average 50%, 80% and 95% coverage for eligible submitted forecasts (S3 Figure 3.2), and average WIS and 95% coverage over time (S3 Figure 3.2). Each figure shows spatial disaggregated results, with increasing population size and quantile numbers. For example, counties with the smallest population are grouped in Quantile 1 and the largest population sizes are grouped in Quantile 5. The following teams are included in these figures: CEID-Walk, LNQ-ens1, Microsoft_DeepSTIA, COVIDhub-4_week_ensemble, COVIDhub-trained_ensemble, COVIDhub-baseline, CU-select, FAIR-NRAR, FRBSF_Wilson-Econometric, IowasStateLW-STEM, JHU_IDD-CovidSP, JHU_CSSE-DECOM, JHUAPL-Bucky, LANL-GrowthRate, LNQ-esn1, UVA-Ensemble.

S3 Figure 3.1. Percent of weeks with complete submissions for all sets of team forecasts, scaled, pairwise relative Weighted Interval Score (rWIS), 95% coverage, and by geographical scale of submitted forecasts. Teams are sorted by increasing rWIS values.

S3 Figure 3.2. Expected and observed coverage rates aggregated over time and horizon for county forecasts. The dashed line represents optimal expected-coverage. Team forecasts that outperformed the COVIDhub-4_week_ensemble model at all coverage levels are labeled on the right hand side of the plots.

S3 Figure 3.3. Mean Weighted Interval Score (WIS) over time, aggregated by geographic units and forecast horizon in A and 95% coverage over time, aggregated by geographic units and forecast horizon in B. The black, dashed vertical line in all panels shows the date that public communication of the case forecasts was paused. The black, dashed horizontal line in panels B show nominal 95% interval coverage

Supporting Information 4: Estimated time-varying reproduction number and epidemic phase classifications. For each state, the top panel shows the median R_t and median upper and lower 90% credible interval over time in red. The bottom panel shows reported case counts over time. Both plots have vertical bands representing the epidemic phase of each forecast week: *increasing*, *peak*, *decreasing*, *nadir.*

Supporting Information 5: Each location specific forecast submitted to the COVID19 Forecast Hub included at least 4 weeks of future predictions. Here, we present disaggregated 1 and 4 week ahead predictions of model performance for each team model that submitted national and state/territory/DC forecasts and were included in the main analyses. Specific plots include the average 50%, 80% and 95% coverage for eligible submitted forecasts (S5 Figure 5.1), average absolute Weighted Interval Score (WIS) and 95% coverage over time (S5 Figure 5.2), and scaled, pairwise rWIS by location (S5 Figure 5.3)

S5 Figure 5.1. Expected and observed coverage rates aggregated for 1 and 4 week ahead forecasts over time for national forecasts in A, state/territory/DC forecasts in B, the largest country forecasts in C. The dashed line represents optimal expected-coverage. Teams that outperformed the COVIDhub-4_week_ensemble model at all coverage levels are labeled on the right-hand side of the plots.

S5 Figure 5.2. Mean Weighted Interval Score (WIS) over time for 1 and 4 week ahead forecasts, aggregated by geographic units, and 95% coverage over time for 1 and 4 week ahead forecasts, aggregated by geographic

units. The black, dashed vertical line in all panels shows the date that public communication of the case forecasts was paused. The black, dashed horizontal line in panels D, E, and F show nominal 95% interval coverage. Teams that submitted national forecasts are presented in A. and D., state/territory/DC forecasts presented in B. and E., and teams that submitted large county level forecasts are presented in C. and F.

S5 Figure 5.3. Scaled, pairwise relative Weighted Interval Score (rWIS; see *Methods* for description) for all teams that submitted national and state/territory/DC forecasts by location for 1 and 4 week ahead horizon. National estimates are displayed first, followed by jurisdictions in alphabetical order. Teams are displayed by decreasing average rWIS across all forecast horizons and locations.

<u>Supporting Information 6</u>*:* Each team model's estimated marginal mean Weighted Interval Score (WIS) over range of reported case counts per epidemic phase. Marginal mean WIS was estimated from GEE model results and reflect values across the 95% confidence interval of mean reported cases. Case counts differ per team model as each team forecasted a different set of locations over a different range of possible dates.

<u>Supporting Information 7</u>: EPIFORGE 2020 guidelines outline 19 recommended reporting items for epidemic forecasting and prediction research (15). These items are included in the checklist below, which also include the page number where each item is described or presented within this evaluation.

Main Text Figures

Figure 1. Weekly incident reported COVID-19 cases per 100K population, nationally (in black) and per state/territory/DC (in gray), over time in panel A. Panel B shows a subset of COVIDhub-4_week_ensemble forecasts (in green) over time, with the median predictions represented as lines and points and the 95% prediction intervals in bands. Reported incident cases (counts per week) are shown in gray. In both plots, the black, dashed vertical line shows the date that public communication of the case forecasts was paused.

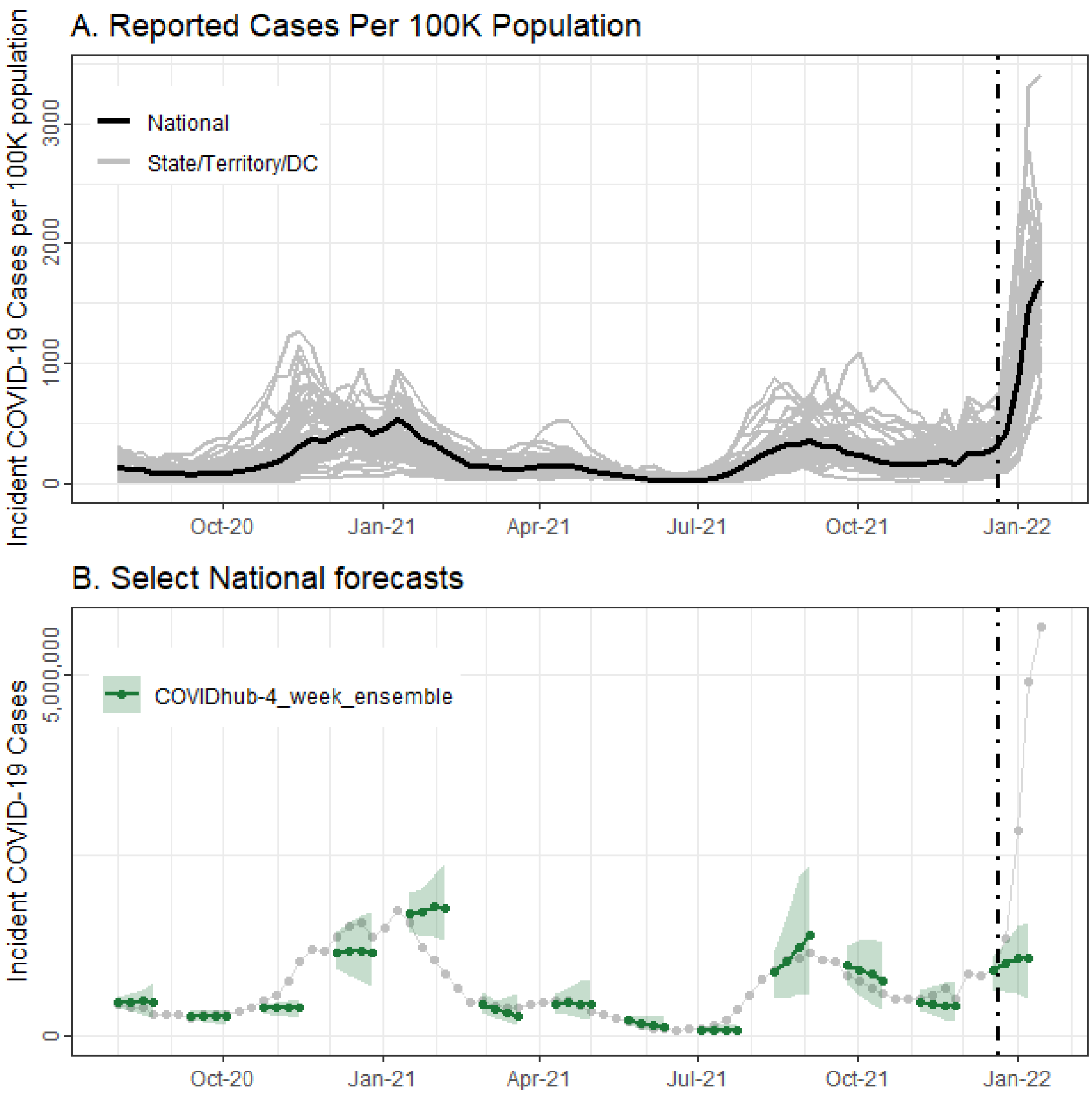


Figure 2: Percent of weeks with complete submissions for all sets of team forecasts, scaled, pairwise relative Weighted Interval Score (rWIS; see *Methods* for description), observed 95% prediction interval coverage, by geographical scale of submitted forecasts. Teams are sorted by increasing state/territory/DC rWIS values.

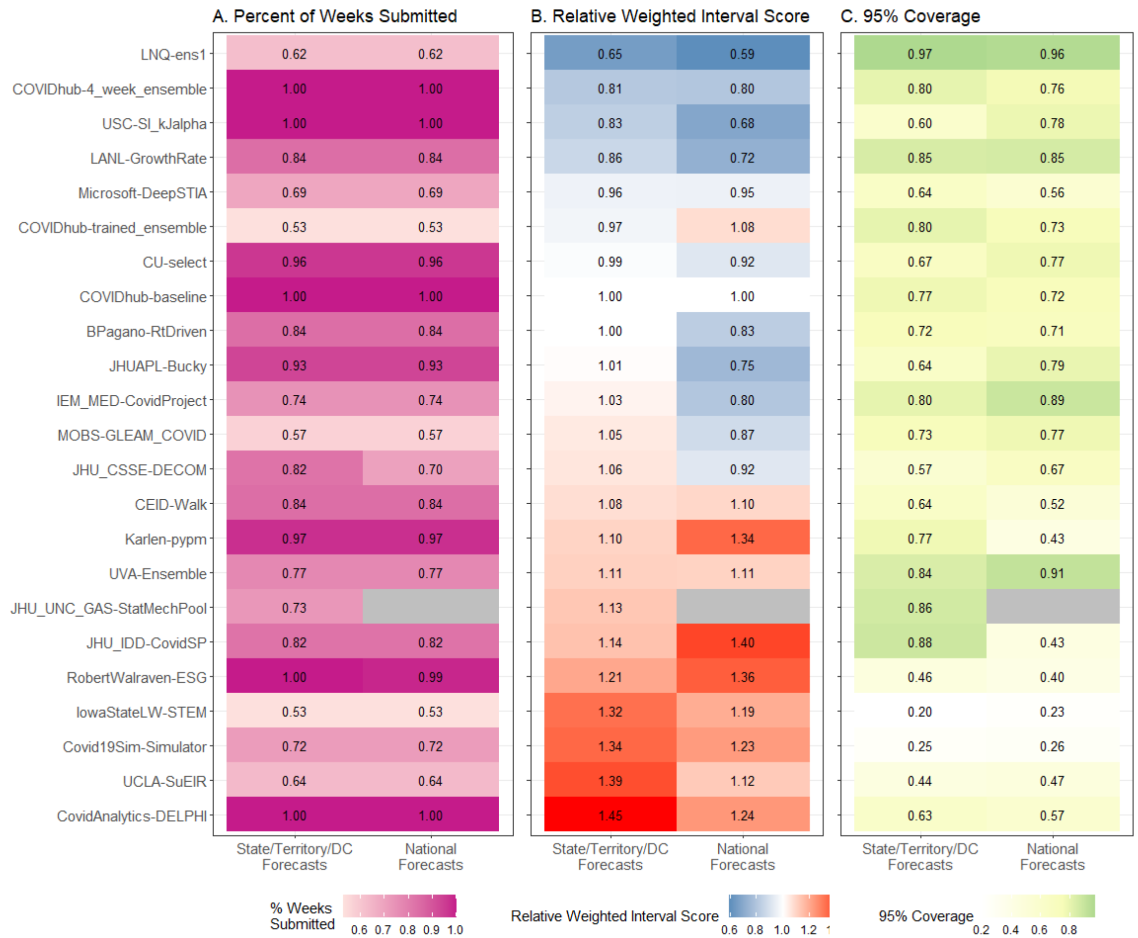


Figure 3: Expected and observed coverage rates for central 50%, 80% and 95% prediction intervals aggregated over time and horizon for national forecasts (panel A), state/territory/DC forecasts (panel B), the largest county forecasts (panel C). The dashed line represents optimal expected-coverage. Team forecasts that had closer to nominal coverage than the COVIDhub-4_week_ensemble model at all three coverage levels are labeled on the right side of the plots.

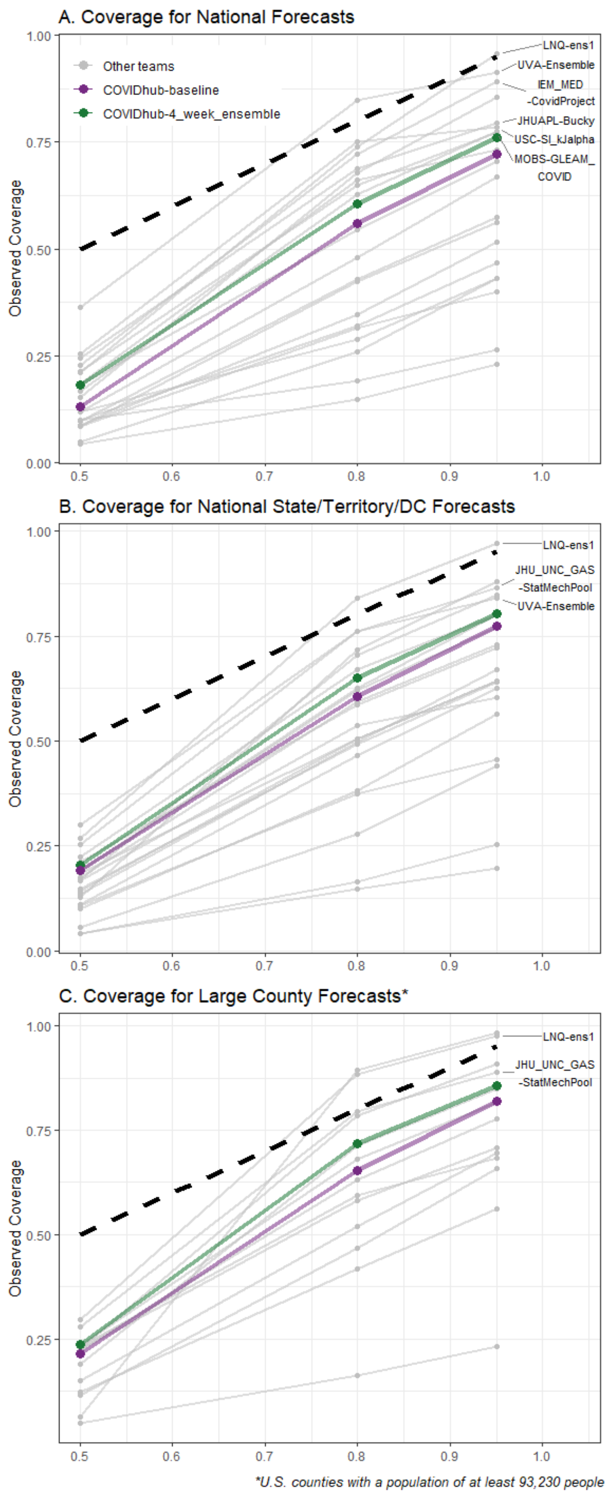


Figure 4: Scaled, pairwise relative Weighted Interval Score (rWIS) (see *Methods* for description) by spatial scale for sets of team forecasts that submitted forecasts for the US nation, states/territories/DC, and all US counties. WIS is averaged across all horizons. The COVIDhub-baseline model has, by definition, a rWIS of 1 (horizontal dashed line). Teams are ordered by increasing state/territory/DC rWIS with the most accurate model on the left. Points for each team are staggered horizontally to show overlapping WIS values.

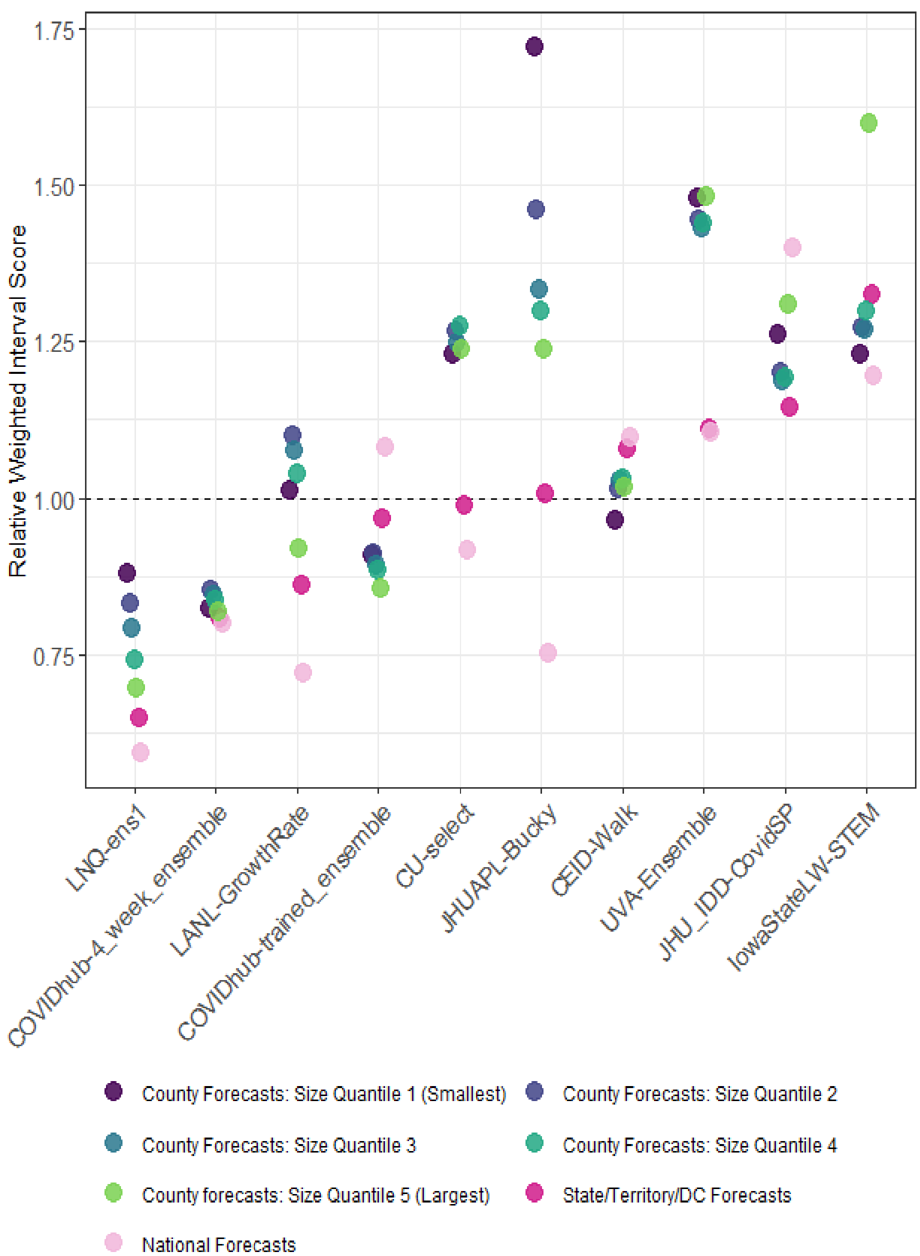


Figure 5: Scaled, pairwise relative Weighted Interval Score (rWIS; see *Methods* for description) by location for national and state/territory/DC forecasts, averaged across all horizons through the entire analysis period. National estimates are displayed first, followed by jurisdictions in alphabetical order. Team forecasts are ordered by increasing average state/territory/DC rWIS.

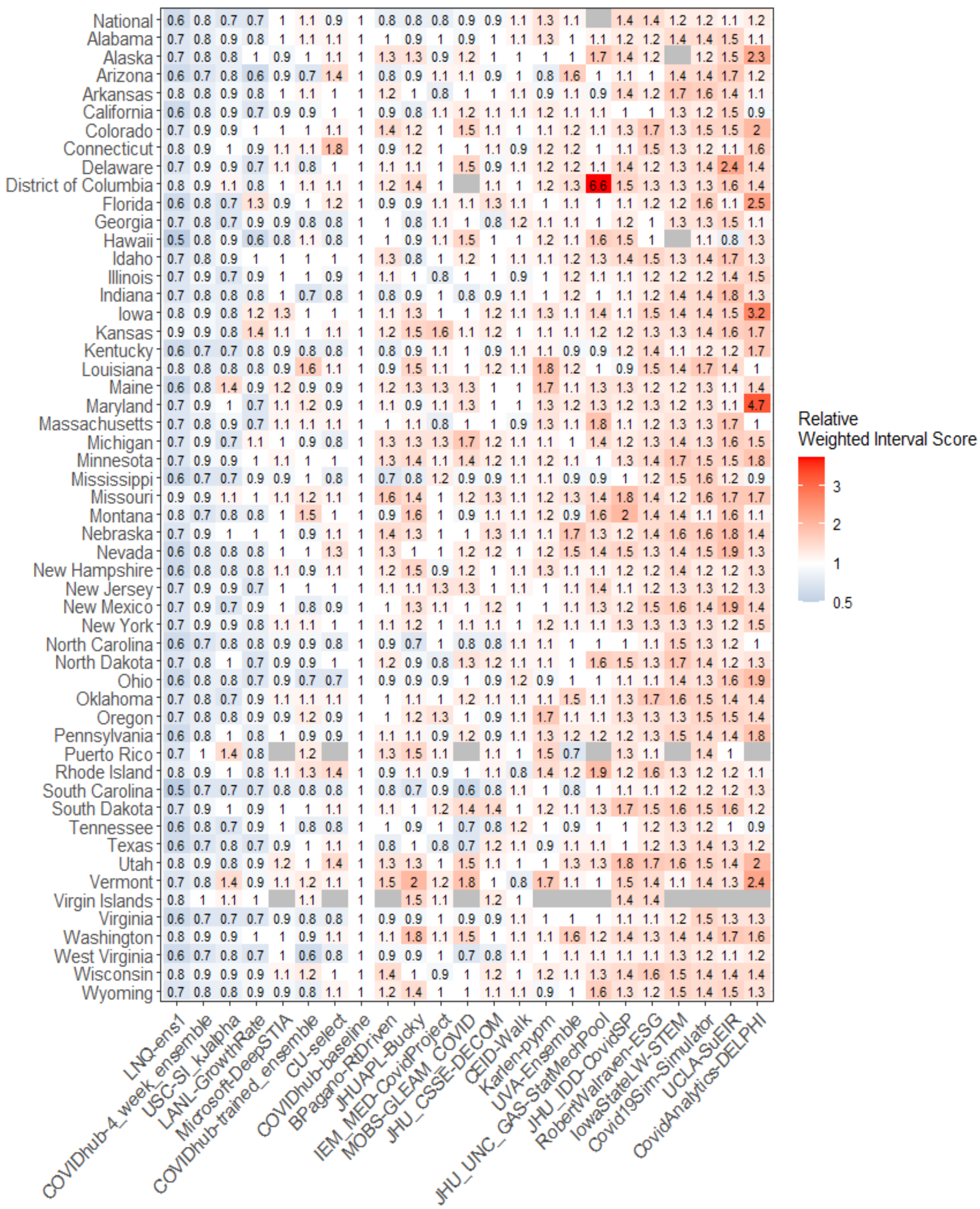
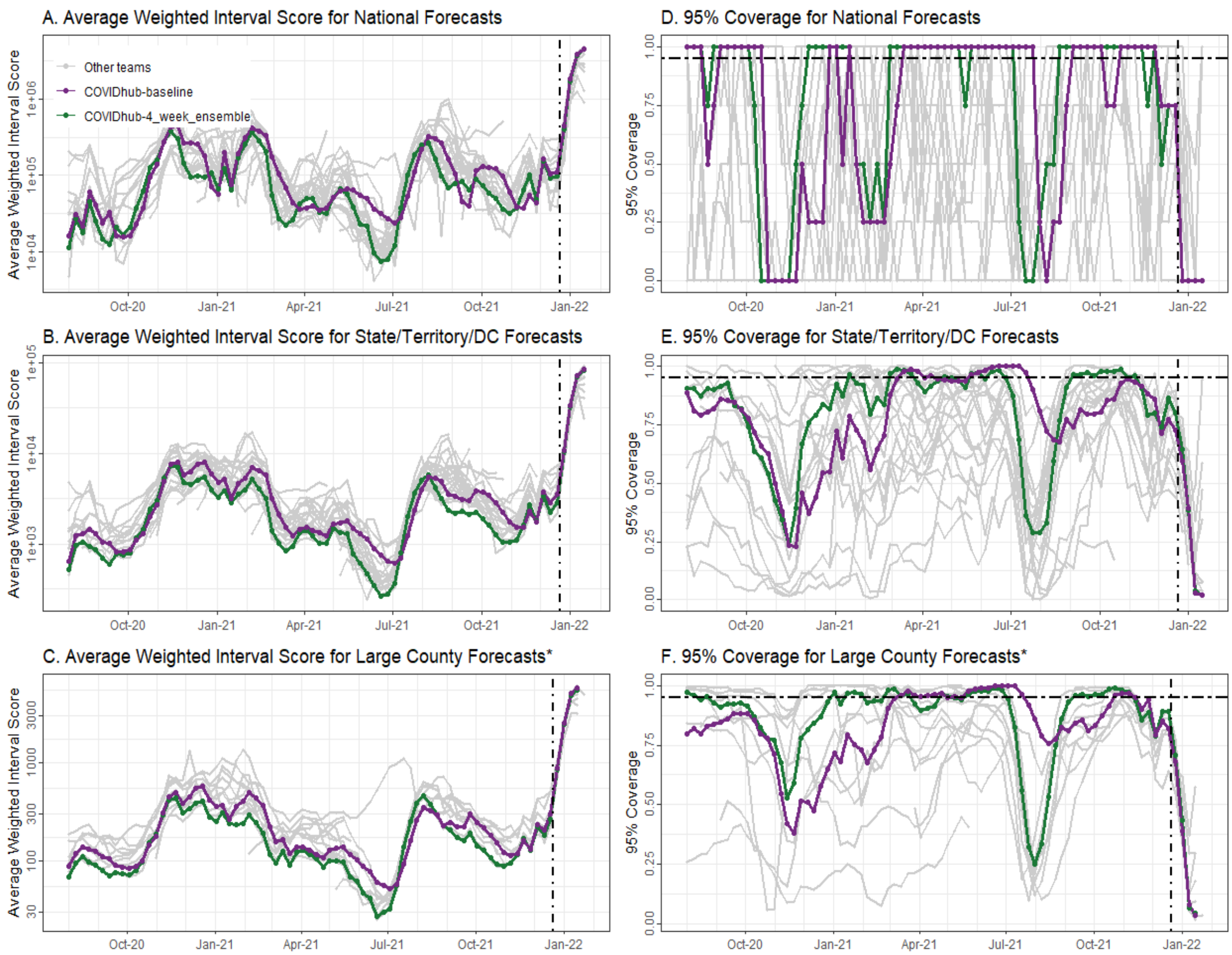


Figure 6: Forecast accuracy over time, aggregated by geographic units, forecast horizon, and prediction date. Panels A-C show average Weighted Interval Score (WIS); panels D-F show 95% prediction interval coverage. The black, dashed vertical line in all panels shows the date that public communication of the case forecasts was paused. The black, dashed horizontal line in panels D-F shows nominal 95% interval coverage. National level forecasts are presented in A and D, state/territory/DC forecasts in B and E and large county level forecasts in C and F.



*U.S. counties with a population of at least 93,230 people

Panels A, D, B and E include: LNQ-ens1, Microsoft_DeepSTIA, COVIDhub-4_week_ensemble, USC-SI_kJalpha, CU-select, LANL-GrowthRate, JHU_CSSE-DECOM, COVIDhub-trained_ensemble, COVIDhub-baseline, Karlen-pypm, BPagano-RtDriven, JHUAPL-Bucky, UVA-Ensemble, IEM_MED-CovidProject, CEID-Walk, Covid19Sim-Simulator, IowasStateLW-STEM, UCLA-SuEIR, JHU_IDD-CovidSP, RobertWalraven-ESG, MOBS-GLEAM_COVID, and CovidAnalytics-DELPHI.

Panels C and F include: LNQ-ens1, COVIDhub-4_week_ensemble, CU-select, LANL-GrowthRate, COVIDhub-trained_ensemble, COVIDhub-baseline, JHUAPL-Bucky, UVA-Ensemble, CEID-Walk, JHU_UNC_GAS-StatMechPoole, IowasStateLW-STEM, JHU_IDD-CovidSP, UMass-MechBayes, FAIR-NRAR, FRBSF_Wilson-Econometric.

Figure 7. Estimated marginal mean Weighted Interval Score (WIS) and 95% confidence intervals for mean cases from team-specific GEE models for all 51 jurisdictions (Panel A). The 95% confidence intervals for the COVIDhub-baseline model are shown in dashed red vertical lines. Panel B presents each team's estimated marginal mean WIS per phase, scaled to the COVIDhub-baseline model's estimated marginal mean WIS for all epidemic phases. Teams with higher estimated marginal mean WIS values (i.e., greater than 1.0) are presented in shades of orange while teams with lower estimated marginal mean WIS (i.e., less than 1.0) are shown in shades of green. Forecasts for a team in a particular phase are marked with an asterisk (*) if the 80% confidence interval of the expected WIS outcome (normalized and on the log scale) was estimated by a model to be lower than the expected WIS of the COVIDhub-baseline model for all phases.